DOCUMENT RESUME

ED 099 427 95 TM 004 305

AUTHOR Knapp, Joan
TITLE A Collection of Criterion-Referenced Tests. TM Report No. 31.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ETS-TM-31
PUB DATE Dec 74
CONTRACT OEC-0-70-3797-519
NOTE 13p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Annotated Bibliographies; *Criterion Referenced Tests; *Elementary Secondary Education; *Test Reviews; *Tests

ABSTRACT

Twenty-one criterion-referenced tests are cited, and for each the following information is provided: description, format and administration, response mode and scoring, technical information, and references. The tests cited are the result of an attempt made to bring together tests designated in the Educational Testing Service Test Collection, a library of tests and test related information, and labeled in the ERIC system as criterion-referenced tests. This annotated bibliography does not list every test that has been labeled criterion-referenced; however, it typifies the variety of tests that are available under the rubric criterion-referenced. Also, criterion-referenced and norm-referenced tests are defined in several ways, and their advantages, limitations, and uses are briefly explored. (RC)

ED ...7

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 31 DECEMBER 1974

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A COLLECTION OF CRITERION-REFERENCED TESTS

Joan Knapp

INTRODUCTION

Criterion-referenced testing has been the focus of a great deal of attention in the measurement community since Glaser (1963) created and defined the term and contrasted it with norm-referenced testing. The criterion-referenced testing movement can, perhaps, be traced back as far as the development of learning theories (Skinner, 1938; Hull, 1945). These theories pointed to the importance of individual differences in learning and encouraged the detailed analysis of both simple and complex tasks. The military put these theories into action during World War II by individualizing instruction through the use of programmed learning experiences. More recently, the development of instructional technology, the legislation of educational accountability, and the existence of such testing-related phenomena as the National Assessment of Educational Progress, with its emphasis on individual exercise reporting, have compelled educators to reexamine existing measurement techniques and to come up with a system of measurement whereby individual progress and mastery can be measured without the use of norming and ranking techniques.

Some observers do not perceive criterion-referenced testing as a new idea. Shub (1973), for example, cites an elaborate civil service examination system that existed in China around 200 B.C. and also calls attention to the tests which Spartan boys had to pass to demonstrate the attainment of the skills of manhood. Ebel (1971) looks to the more recent past and recalls the percent grades that were used almost universally in schools and colleges in the 1920s—grades that indicated, however imprecisely, the amount of material students had mastered. Block (1971) reviews a number of these previous attempts at mastery learning and diagnosis in education.

Although most educators would agree that it would be extremely useful to know exactly what a student knows or can do, measurement specialists do not agree either on the approaches that would be most likely to lead to this goal or on the extent to which the goal is attainable. A number of attempts have been made to integrate explorations of criterion referencing with such previously examined measurement concepts as diagnostic, subject-mastery, performance, and minimum-competency testing. Much development work has centered on measurement of psychomotor skills, job-related performance, and introductory mathematics and reading skills, but there has been only limited application of criterion-referenced techniques to such complex operations, for example, as verbal comprehension.

What Is a Criterion-Referenced Test?

Interestingly, a generally accepted definition for the term *norm-referenced test* (NRT) can easily be found in educational measurement literature, but varying definitions of the term *criterion-referenced test* (CRT) exist. A norm-referenced test is usually defined as a test which gives us information that compares an individual's performance with that of others. Scores on a test or set of tasks are related (referenced) to a table of norms.

There is disagreement as to what is meant by a criterion-referenced test (CRT). Glaser and Nitko (1971) define a criterion-referenced test as "a test that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." Ivens (1972) states that two characteristics defining criterion-referenced measurement are "the presence of a performance criterion, and test items keyed to a set of behavioral objectives." In a more specific statement, Jackson (1970) applies the term *criterion-referenced* only to a test "*designed* and *constructed* in a manner that defines explicit rules linking patterns of test performance to a behavioral reference." Livingston (1972) refers to a criterion-referenced test as "any test for which a criterion score is specified *without* reference to the distribution of scores of a group of examinees."

TM 004 205

This publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

Kriewall (1972) states that CRTs are content-specific measures that can be useful for the stratification of learning groups on a day-to-day basis if need be.

Still another definition is posited by Fremer (1972), who perceives a clear conceptual bridge between NRTs and CRTs and notes that for group measurement any test may be criterion-referenced if test scores are related to levels of performance on some behavioral scale. Using a model of correlational consistency, he suggests that tables of relationship be developed that link test-score levels to behavioral-criterion levels.

Uses of CRTs and NRTs

Unfortunately, because of the recent surge of interest in CRTs, NRTs are often cited as inapplicable or inappropriate to assessment and evaluation in education. Researchers frequently point to the disadvantages of NRTs and to what they *cannot* do (Housden and LeGear, in press). NRTs have even been described as detrimental to the educational and sociological development of individuals because they "provide a situation where half of the American school children must always be below the median (or always 'lose'), thus linking scores on an NRT to the notion of a self-fulfilling prophecy" (Brazziel, 1972). Brazziel goes one step further and asserts that the advantages of CRTs far outweigh the disadvantages. An extreme view is taken by Gentile (1971), who states that NRTs are "sadistic, unethical, and statistically unsound."

It is clear that there are situations for which NRTs are appropriate and others for which CRTs are indicated. Garvin (1970), for example, points out that researchers should be deciding which situations call for CRTs and which for NRTs rather than taking a stand regarding the general appropriateness of each approach to testing. He lists several situations that require CRTs, including those in which specific units of instruction are primarily intended to prepare individuals for the next unit in a sequence. He asserts that when a predetermined performance level in certain tasks is crucial CRTs should be used.

Klein (1972) indicates that both CRTs and NRTs should be based on objectives and that the problem with many nationally normed tests is that they are not constructed with objectives in mind. Further, Klein looks at the diverse purposes of educational evaluation and claims that the difference between NRTs and CRTs rests in the interpretation of the results of the two kinds of tests. Cox (1970) concurs. Using a graduate measurement examination as an example, he illustrates how it is possible for a test to yield both norm-referenced and criterion-referenced information. Because of this, he concludes that a CRT need not be operationally defined in such restricted terms.

Limitations and Problems in Criterion-Referenced Testing

Ebel (1971) summarizes some of the limitations of CRTs:

1. They do not tell us all we need to know about achievement. [A CRT is a coarse test—a pass-fail kind of situation.]
2. Detailed specification of educational objectives is not essential to effective education. [The drawing up of specific, detailed objectives costs more in time and effort than they are worth and such objectives are more likely to suppress than to stimulate effective thinking.]
3. CRTs are difficult to construct on any sound basis. [These tests are difficult to validate. It is difficult to specify clearly what criterion is being used.]

Hastily conceived criterion-referenced testing programs that use items which are loosely connected to learning objectives can do a great deal of harm to instructors and students (Hawes, 1973). Further, there are many commercially available CRTs with virtually no information accompanying them. It would appear that as long as a test is dubbed criterion-referenced by the publisher, it is thought to be one no matter how it is constructed and no matter whether any data is available to the user for the interpretation of test results.

Fitzgibbon (1972) warns that test publishers must take an ethical stance when using criterion-referenced testing in statewide assessment programs.

> Statewide assessment programs are conceived in an environment of suspicion, gestated during a period when legislators, school administrators and teachers are increasingly hostile to one another. Finally, tests must be delivered because of unreal time constraints.

He suggests that publishers should not bid their services when a criterion-referenced testing project seems unrealistic and impossible.

COLLECTION OF CRITERION-REFERENCED TESTS

The following collection represents an attempt to bring together tests designated in the Test Collection office of Educational Testing Service and labeled in the ERIC system as *criterion-referenced tests.* The collection does not contain every test that has been labeled *criterion-referenced*; however, it typifies the variety of tests that are available under the rubric *criterion-referenced.* Tests which had little information, data, or description associated with them were not included. Test and instructional objective systems such as SOBAR and IOX are described by Klein and Kosecoff (1973). Information on district and statewide criterion-referenced testing programs can be obtained from individual state departments of education.

The APELL Test by Eleanor V. Cochran and James L. Shannon. © 1969, EDCODYNE Corporation.

Description: A system of instructional diagnosis and design which assists the teacher in assessing the child's school entry-level development, the test diagnoses skill levels based on specific performance objectives covering: Pre-Reading (visual discrimination, auditory discrimination, letter names); Pre-Math (attributes, number concepts, number facts); and Language (nouns, pronouns, verbs, adjectives, plurals, prepositions). The manual gives a rationale for each objective and suggests related instructional activities. A Spanish translation of the test is available. The 50-item pictorial instrument is designed for children age 4-1/2 to 7, but it may be useful for somewhat younger or older children. A supplement to the manual expresses the hope that the test be used "to support the empirical notion of performance rather than the usual norm comparisons associated with standardized tests."

Format and Administration: The questions are administered orally to individuals or to a small group. The test booklet, designed to present only one item at a time by flipping multi-colored cards, presents three options for each question by means of pictures. The test takes about 40 minutes and should be administered in two sittings.

Response Mode and Scoring: The child either says the answer or points to the picture which he/she thinks answers the question. Answer or machine-scorable cards are scored by the EDCODYNE Corporation. Scoring services include: a total score and subscore for each pupil; adhesive-backed student profile forms; class profile; calculation of the standard deviation for each subtest and for the total test.

Technical Information: The test was developed using a sample of 5,000 children (15 percent black; 23 percent Spanish surname; 7 percent oriental, American Indian and other subgroups). The test-retest reliability coefficient using a sample of 85, with a two-week interval, was .81. The correlation coefficient between APELL scores and Metropolitan scores was .78 (n=104).

Review:
Proger, B. *The Journal of Special Education,* 1971, 5, 195-198.

Basic Concept Inventory (field research edition) by Siegfried Engelman. © 1967, Follett Educational Corporation.

Description: The author intends the inventory to be used primarily for culturally disadvantaged preschool and kindergarten children, slow learners, emotionally disturbed children, and mentally retarded children. It is a broad checklist of basic concepts that are involved in new learning situations. The inventory may be used for diagnostic purposes or for group placement. Subjectively selected subtests are: Basic Concepts; Statement Repetition and Comprehension; and Pattern Awareness. The test is described as a criterion-referenced test in that it was designed "to evaluate the instruction the child has received on specific, relevant skills."

Format and Administration: The individually administered inventory consists of nine stimulus cards about which the administrator asks questions. The number of questions per card range from 1 to 10. The procedure takes 15-20 minutes. The examiner may be a classroom teacher or someone with special training depending on the nature of the diagnostic and/or remedial decisions to be made from test scores. Special instructions are provided for administering the test to handicapped children.

Response Mode and Scoring: Responses are made verbally or the child points to the object in the picture which he or she thinks answers the question. The examiner records the responses on an answer sheet and then scores the test using the manual, which gives guidelines for scoring and examples of case histories.

Technical Information: Though the manual gives some sample data, there are no norms. No information on reliability or validity was available in 1967; however, the 1967 edition states that the reliability data is to be available in 1968 or 1969. The author outlines his procedures for assuring content and criterial validity (validity related to criterion-referenced measurement).

References:

Buros, O.K. (Ed.) *Seventh mental measurements yearbook. Vol. II.* Highland Park, New Jersey: The Gryphon Press, 1965.

Hofmeister, A., & Espeseth, V. K. Predicting academic achievement with TMR adults and teenagers. *American Journal of Mental Deficiency*, 1970, 75, 105-107.

Sears, C.R. A comparison of the basic language concepts and psycholinguistic abilities of second grade boys who demonstrate average and below-average levels of reading achievement. Unpublished doctoral dissertation, Colorado State College, 1969.

Experiences with Sets and Numbers: Mathematics Evaluation Materials Package Project by Howard Russell. © 1972. The Ontario Institute for Studies in Education, 252 Bloor Street West, Toronto 5, Ontario, Canada.

Description: The Mathematics Evaluation Materials Package (MEMP) is a set of performance objectives and companion test items for mathematics education in grades 4 to 6. "Experiences with Sets and Numbers" includes items on: place values; Roman numerals; ordering of whole numbers: >, < and =; special subsets of the whole numbers; intersection and union of sets; rounding numbers; and positive and negative numbers in concrete situations.

Format and Administration: The teacher selects from the package those items he/she feels are appropriate for his/her curriculum. MEMP may be used to design tests covering short units of work or to construct longer tests. Sample inventories are included in the package.

Response Mode and Scoring: Answers to the test items are provided.

Technical Information: No technical information is given.

References: None cited by author.

Explorations in Biology by Eugenia M. Koos and James Y. Chan. © 1972. Mid-Continent Regional Educational Laboratory.

Description: This is a series of eight parallel, single-topic tests designed to measure the attainment of 14 objectives concerned with inquiry skills in biology. The single-topic, simulation format was selected to accommodate the unitary nature of an inquiry. Objectives selected for the tests were based on studies by various science educators and curriculum specialists. These objectives were also in agreement with the goals and outcomes of the BSCS Biology curriculum. The series is intended for the average sophomore student in the first course in high school biology. Some of the target group taking the tests as a pretest in the fall may demonstrate attainment of a particular criterion level. There would then be no need for these particular students to take instruction intended to guide the target group toward this level. The items were written by educators and test construction specialists familiar with tenth-grade biology curriculum and with inquiry processes.

Format and Administration: The number of response options (2 to 5) offered for each item varies from section to section. All questions are read by the students and are answered on an answer sheet. The tests may be administered by a regular classroom teacher.

Response Mode and Scoring: The examinee blackens the letter or number of the answer chosen on the answer sheet. Because the scoring of responses is based on the judgment of test writers as to the preferred inquiry process response and not on answers that are related to memory for facts or understanding of concepts, a weighted scoring system is used. This procedure results in some variability among the total scores possible for the various booklets.

Technical Information: The authors used the *Behavioral Checklist for Science Students* and total EIB scores to investigate the relationship between ratings by peers and by teachers. There did not appear to be a linear relationship between the two sets of scores. When curvilinear techniques were applied to the data, the correlation was .36 with an N of 150. Scores on the *BSCS Processes of Science Test* were correlated with scores on the EIB battery, resulting in a nonsignificant coefficient. To assure content validity, persons well acquainted with inquiry instruction were asked to judge whether EIB items could be referenced to the Mid-Continent Lab's BSCS objectives. With the exception of one judge and one section of items, all items were referenced appropriately. Using construct validation, the authors correlated scores on the EIB with scores on 10 aptitude and achievement tests. These tests included the SCAT and DAT. The results suggest that the skills assessed by the EIB are rather highly related to verbal reasoning and to a lesser extent to critical thinking. The initial index of internal consistency was .96 (N=451). Another estimate of .74 (N=150) was obtained in a more homogeneous sampling. A study was made of the test-retest stability of the EIB. This yielded a coefficient of .66 (N=52) with one week intervening between test administrations. However, this coefficient does not appear to be appropriate for a sequential test such as the EIB. Furthermore, the probability was great that the student's second response to the instrument would be altered because of his exposure to readings contained in the test.

References:
Koos, E.M., & Chan, J.Y. Criterion-referenced tests in biology. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

FLES Criterion Referenced Tests by Robert M. Offenberg, David Montalvo, and Edward K. Brown. © 1971, Office of Research and Evaluation, School District of Philadelphia.

Description: Three criterion-referenced tests were developed to evaluate the Latin FLES program in the School District of Philadelphia. The testing involved 4,000 fourth-, fifth-, and sixth-grade pupils receiving 15 to 20 minutes Latin instruction daily from Latin teachers who served several schools. The primary objectives of the program were: 1) to introduce children to basic Latin structure and vocabulary; 2) to extend the English vocabulary of children through the study of Latin roots and affixes; and 3) to acquaint children with classical culture and its influence on the present. Achievement was measured by means of a criterion-referenced test battery containing a *Latin Culture Test*, a *Word Power Game*, and an *Oral Latin Test*. The *Latin Culture Test* consists of two parts. Items on the test are taken from the fifth-grade course of study and constitute a sample of major facts and concepts which the pupils should have acquired if they succeeded in mastering the culture curriculum for the first year of study. The test was designed so that students with adequate mastery should score at least 75 percent and students doing minimally passing work should score about 60 percent. The *Word Power Game* was designed to assess student mastery of English skills. It is divided into three parts, the first of which contains nine items which check the student's knowledge of English derivatives and cognates actually appearing in the program. Items in the next part are based on English vocabulary not taught in the program but derivable from Latin roots and affixes. The third section contains five items which are based on the material included in the program but which differ from the first two parts (administered orally) in that the pupil is required to read the item. The *Oral Latin Test* contains material taken directly from the course of study or the visual aids (flash cards and so forth) designed for use with it.

Format and Administration: The *Latin Culture Test* takes a maximum of 20 minutes of testing time. Questions and four choices for answers are presented orally to the students in a group. The *Word Power Game* is administered in a similar fashion. Testing time is not indicated. The *Oral Latin Test* is administered individually. Testing time is not given.

Response Mode and Scoring: With the exception of the oral test, students circle the letters of the answers they choose. Answers are keypunched and analyzed by means of the QUIKSCORE program owned by the University of Pennsylvania. The *Oral Latin Test* is scored by the Latin teacher, who decides whether the pupil's answer is correct or incorrect. At the end of the test, the teacher writes down the number of correct answers to the 10 questions.

Technical Information: Other instruments were included in the evaluation. They were: (1) principal's questionnaire; (2) cooperating teacher's evaluation form; (3) pupil's questionnaire; (4) parents' questionnaire; and (5) Iowa Tests of Basic Skills, Vocabulary (V) subscale. Difficulty indices are given for individual items and/or groups of items. No information is given concerning reliability indices. When pupil performance on the Iowa subtest was examined, it was found that FLES program pupils were functioning about on grade level, whereas a control group of pupils were functioning one year below grade level.

References:
Offenberg, R.M., Montalvo, D., & Brown, E.K. Evaluation of the elementary school (FLES) Latin program 1970-71. Report No. 7202, Office of Research and Evaluation, School District of Philadelphia, October, 1971. ED 056 612.

Illinois Tests in the Teaching of English by William H. Evans and Paul H. Jacobs. © 1969, 1972, Southern Illinois University Press.

Description: This is a battery of four tests designed to measure the achievement of preservice and inservice high school English teachers on the basis of certain objectives established by specialists in education and by practicing teachers. The authors describe the tests as criterion-referenced in that their purpose is "to measure the individual teacher's achievement of certain objectives based on criteria established by his/her colleagues." *Knowledge of Language: Competency Test A* covers (1) the functioning of language; (2) the principles of semantics; (3) systems of English grammar; (4) the history of the English language, including its phonological, morphological, and syntactic changes; and (5) concepts about levels of usage and dialectology (84 items). *Attitude and Knowledge in Written Composition: Competency Test B* is concerned with teachers' attitudes toward or philosophy of the teaching of written composition. It also assesses their ability to recognize characteristics of good writing, perceive the complexities of composing, and recognize and analyze the strengths and weaknesses of a composition and communicate this analysis effectively (62 items). *Knowledge of Literature: Competency Test C* is designed

to assess teachers' familiarity with (1) patterns of development of English and American literature; (2) major authors representative of various genres and periods; (3) literature concerning minority groups; (4) both ancient and modern major works of literature; (5) major critical theories and schools of criticism; and (6) literature suitable for adolescents. It also tests their ability to closely read literary texts (143 items total). *Knowledge of the Teaching of English: Competency Test D* assesses teachers' knowledge of (1) learning processes and adolescent psychology; (2) corrective and developmental reading techniques; (3) principles of evaluation and test contruction; (4) concepts of the role of English in the total secondary school program; (5) principles of curriculum development; and (6) ways to select and adapt materials and to guide, stimulate and challenge students (104 items).

Format and Administration: The tests all consist of multiple-choice items and are administered to preservice or inservice high school English teachers. The tests are not timed, but tests A and B should each take 45 minutes to an hour, while tests C and D may require more time.

Response Mode and Scoring: Examinees use a machine-scorable answer sheet; however, a hand-scoring matrix is available. The Southern Illinois University Testing Center provides machine-scoring services. The basic scoring service provides a roster showing each person's performance on each test. Optional services include: item analysis, individual profile reports, punched cards, and magnetic tape.

Technical Information: The test administrator's manual gives the results of field tests of the experimental edition (i.e., item analysis, ranges, reliability estimates); however, the results of operational or research testing are not reported. Because the tests are criterion-referenced the authors did not establish national or regional norms; however, teacher-training institutions and school systems can develop local norms on A, C, and D if they wish to.

References: None cited by authors.

The Instant Word Recognition Test by Edward Fry. © 1971, Educational Systems, Inc.

Description: This test measures sight recognition of the Instant Words in order to determine the starting point in teaching Instant Words, a graded high-frequency reading vocabulary. It may also be used to determine general reading achievement for group placement. It is available in two equivalent forms for each of two levels.

Format and Administration: The test is intended for grades 1 through 4. It may also be administered to older children in remedial reading programs. It can be administered individually or to small groups by a teacher without instruction. The examiner reads 48 lines of words and sentences containing the keyed word. The answer sheet contains five words for each sentence, one of which is the word the examiner has read. No time estimate is given.

Response Mode and Scoring: The student selects from the five words the word that he thinks the examiner has read by putting an *X* on the word. The tests are self-scoring in that a special carboned sheet is attached to the student's answer sheet.

Technical Information: The author states that "though this test does not have a large standardization group, it has been administered to 153 first graders and their mean score was 11.11." This same sample was given the paragraph meaning subtest of the Stanford Achievement Test. The correlation coefficient between the two tests was 0.79.

References: None cited by author.

KeyMath Diagnostic Arithmetic Test by Austin J. Connolly, William Nachtman, and Mil Prichett. © 1971, American Guidance Service, Inc.

Description: This diagnostic test of mathematics skills covers: Content (numerations, fractions, geometry, and symbols); Operations (addition, subtraction, multiplication, division, mental computation, and numerical reasoning); and Applications (word problems, missing elements, money, measurement, and time). The items in each subtest are arranged in order of difficulty. The test manual lists a behavioral objective for each item. Successful performance on any item implies mastery of the skill sampled at that level. "The identification of these abilities in behavioral terms and the identification of their relative level of difficulty provide the examiner with a criterion-referenced scale."

Format and Administration: The test is individually administered to children in preschool through grade 6 and consists of full color illustrations on cards set up in a book like an easel. There is no upper-grade limit for remedial use. The examiner need only be familiar with the test manual and the topics the test covers. The KeyMath kit includes 209 items. Only those items appropriate for the student's level of achievement are administered. Although the test is not timed, it should take about 30 minutes.

Response Mode and Scoring: Most items require the subjects to respond verbally to open-ended questions that are presented orally by the examiner. The test is scored by the examiner. The manual gives guidelines for interpreting the 14 subscores, but interpretation of test results is most informatively done by a professional.

Technical Information: The authors have provided grade norms (norming sample — n = 1,222) even though they feel that KeyMath is a diagnostic instrument and should be used to provide information about the student's individual performance on subtests and particular items. Internal consistency-reliability coefficients were calculated for subtests, total test, and for samples which ranged from grades K to 7. The median reliability coefficient for total test score across grades is .96. Median reliabilities on subtests ranged from .64 to .84. Content validity was assured by expressing each item in behavioral terms in the form of objectives. Scores on a pretest of KeyMath correlated with scores on the arithmetic portion of the Iowa Test of Basic Skills resulted in r .69. Research is now being undertaken on the final instrument to add to the construct validity of the test.

Review:

Bannatyne, A. *Journal of Learning Disabilities*, 1973, 6 (3), 131-132. (See test manual for an additional list of references and research studies in which KeyMath was used.)

Ohio Vocational Interest Survey (OVIS) by Ayres G. D'Costa, David W. Winefordner, John G. Odgers, and Paul B. Koons, Jr. © 1969, Harcourt Brace Jovanovich, Inc.

Description: This survey is designed to help students in grades 8 to 12 with educational and vocational planning. Scores rank students' vocational interests along 24 scales derived from the Dictionary of Occupational Titles (DOT), a classification of 21,741 jobs in terms of involvement with people, data, and things.

Format and Administration: The survey has two parts. In the first, the student indicates his preference among 27 general types of jobs. In the second, the student indicates how much he would like each of 280 specific kinds of work. The school administering the survey has the option of adding eight questions to the survey to obtain additional information for counseling or curriculum planning. Administering the survey requires 60 to 90 minutes.

Response Mode and Scoring: The survey is available in two editions, machine scorable booklets and reusable booklets with machine scorable answer sheets. The basic scoring service provided by Harcourt Brace Jovanovich, Inc., records each student's responses to the questions in the first part and reports his interests as indicated by the second part in terms of raw scores, percentile, and stanine on each of the 24 scales. An indication of response consistency is also provided. Optional scoring services include: a group summary report, permanent record labels, IBM work cards, and computer tapes. Guidelines for interpreting survey results are given in the manual.

Technical Information: Scoring may be done with reference to national norms, local norms, or both. The manual appendix includes summary reports on the norming sample, scale intercorrelations, a description of each item in terms of the DOT classification, and test-retest reliability and SE_m data for grades 8 and 10. Test-retest reliability ranges from .775 to .848. Standards of item performance and scale homogeneity are included in the discussion of validity.

Review:

Taylor, R. *Journal of Educational Measurement*, 1972, 9, 88-91.

Oral Reading Criterion Test for Determining Independent and Instructional Reading Levels by Edward [illegible] © 1971, Dreier Educational Systems, Inc.

Description: The test was devised to determine the independent and instructional reading levels of both children and adults. Simple first-grade primer material is the lowest level on the test. The highest level, the seventh-grade reading level, is representative of popular adult and nontechnical secondary reading levels. A chart on the last page of the test enables the examiner to determine the difficulty level of instructional materials that the teacher plans to use.

Format and Administration: The test is individually administered to children in grades 1 to 7 or to adults who need instruction in reading. It consists of 10 paragraphs, covering the range of difficulty levels, to be read aloud. The paragraphs are not timed.

Response Mode and Scoring: The student reads the paragraphs aloud while the examiner scores the test on the basis of misreadings, stumbling, and omitted words. Scoring information for the independent, instructional, and frustration levels are given on the test.

Technical Information: No technical information is given.

References:

The author cites the April 1968 *Journal of Reading* and the March 1969 *Reading Teacher* but does not give titles of articles or the page numbers.

Phonics Criterion Test of 99 Phoneme Grapheme Correspondences by Edward Fry, © 1971, Dreier Educational Systems, Inc.

Description: This test uses nonsense syllables to determine areas of difficulty in phonics. The test covers: easy consonants, short vowels, long and silent vowels, difficult consonants, consonant digraphs, consonant second sounds, schwa sounds, long vowel digraphs, vowel plus *r*,

broad *o*, diphthongs, difficult vowels, consonant blends, and consonant exceptions.

Format and Administration: The test is individually administered to children in grades 1-3. There are 99 items. No time estimate for administration is given.

Response Mode and Scoring: The children are asked to decide how nonsense syllables are pronounced, and the test is scored by the examiner.

Technical Information: No technical information is given.

References: None cited by author.

Prescriptive Mathematics Inventory by John Gessell. © 1972. CTB McGraw-Hill.

Description: This test assesses mastery of objectives normally covered in grades 4 through 8 mathematics curricula. The test may be used for planning individual, small group, or classroom instruction and for assessing, after a period of instruction, what progress the students have made. References to mathematics materials and learning activities corresponding to specific objectives are provided. The test is available in four separate but overlapping levels. The items in the *Orange Book* deal with the addition, subtraction, multiplication, and division of whole numbers; with the properties of these operations; with number theory, measurement, non-metric geometry, place value, and problem solving. In addition to the above, the *Aqua Book* covers problem solving and basic operations with fractions and decimals. The *Purple Book* does not include place value problems. It adds problems on numeration systems, percent, sets, and statistics. Level C has fewer items on basic operations than the lower levels. It does test operations with negative numbers and it adds functions probability, trigonometry, and reasoning to the topics covered in the *Purple Book*. Interim Evaluation Tests for use on completion of each unit of study have been developed for all but the C level.

Format and Administration: The test is administered to groups. The *Orange Book*, covering material usually taught in grades 4 and 5, requires about 3-1/4 hours. The *Aqua Book*, which covers objectives taught in grades 5 and 6, requires about 4 hours. Both the *Purple Book*, which covers materials taught in grades 6 and 7, and Level C, which covers material taught in grades 7 and 8, require about 4-3/4 hours. Test administration must be divided into sessions not exceeding one hour in length.

Scoring: The students code their answers to the open-ended questions on a special machine-scorable grid, and the test is machine scored by the publisher. Reporting services include: an Individual Diagnostic Matrix for each student; a Class Diagnostic Matrix; a Class Grouping Report; an Individual Study Guide that refers each student to pertinent pages in the text the class is using; a Master Reference Guide that refers the teacher to the class text pages that teach, review, or test the objectives measured by the inventory.

Technical Information: No technical information is given.

References:

Roudabush, G.E., & Green, D.R. Some reliability problems in a criterion-referenced test. Paper presented at annual meeting of the American Educational Research Association, New York, 1971.

Roudabush, G.E., & Green, D.R. Aspects of a methodology for creating criterion-referenced tests. Paper presented at annual meeting of National Council for Measurement in Education, Chicago, 1972.

Prescriptive Reading Inventory by Elizabeth M. Layman. © 1972, CTB/McGraw Hill.

Description: This test is designed to provide information useful in planning individual and class reading instruction. Behavioral objectives are grouped into seven areas: Recognition of Sound and Symbol; Phonic Analysis; Structural Analysis; Translation; Literal Comprehension; Interpretive Comprehension; and Critical Comprehension. The inventory is available in four separate but overlapping levels: the *Red Book* (Level A), the *Green Book* (Level B), the *Blue Book* (Level C), and the *Orange Book* (Level D). Suggested classroom activities and references to textbook series and reading programs are provided for all the objectives. The publisher states that the PRI is a criterion-referenced test "designed to provide evaluation relevant to classroom instruction; that is PRI evaluates each student's mastery of specific behavioral objectives."

Format and Administration: The questions in levels A and B are presented orally to a group of pupils; in levels C and D, the stimulus material is written. Level A, appropriate for grades 1-1/2 to 2-1/2, requires about three hours of actual testing time. Both Level B (grades 2 to 3-1/2) and Level C (grades 3 to 4-1/2) require about 2-3/4 hours. Level D (grades 4 to 6-1/2) requires about 2-1/2 hours. Testing should be divided into sessions of not more than 45 minutes each.

Response Mode and Scoring: Students use machine-scorable answer sheets, and the test is machine scored by the publisher. Reporting services consist of: an Individual Diagnostic Map and an Individual Study Guide for each pupil; a Class Diagnostic Map and a Class Grouping Report; and a Program Reference Guide, which refers the teacher to the pages in the reading pro-

gran the class is using which are pertinent to each PRI objective.

Technical Information: The development of the tests is described in the manual. A Technical Report is scheduled for publication in 1974.

References: None cited by publisher.

The Progressive Achievement Tests by Warwick B. Elley and Neil A. Reid. © 1969, 1970, 1971, Whitcombe & Tombs, Ltd.

Description: This battery represents three language skills tests developed for use in the New Zealand schools. The *Reading Comprehension* test (©1969) assesses factual and inferential comprehension of prose material. The *Reading Vocabulary* test (©1969) provides an estimate of what proportion of the Wright list of the 10,000 most common words the student knows. The *Listening Comprehension* test (©1970) assesses both simple recall skills (receptive listening) and inferential skills (reflective listening); it may be used to identify children who need help with listening skills, as an estimate of "reading expectancy," or simply as an additional measure of verbal skill. The *Reading Comprehension* test and *Reading Vocabulary* are available in three forms, A, B, and C, with form C reserved for research purposes. The *Listening Comprehension* test is available in two forms, A and B. Elley outlines a procedure whereby content-referenced scores can be developed for use with these tests (see reference).

Format and Administration: All three tests are multiple-choice and may be administered by a regular classroom teacher. The tests are appropriate for students ages 8 to 15. The *Listening Comprehension* test may also be useful for 7 or 16 year-olds. Students in different grades take different but overlapping parts of each test. The *Reading Comprehension* test takes 40 minutes, the *Reading Vocabulary* 30 minutes, and the *Listening Comprehension* about 50 minutes.

Response Mode and Scoring: The examinee writes the letter of the answer chosen in the appropriate space on the answer sheet. Answer sheets are hand-scored by the teacher using a stencil key provided.

Technical Information: Raw scores may be converted into grade-level scores or percentiles. Further technical information given in the manual includes the split-half reliabilities of each form of each test at each grade level. Reliability coefficients range from .88 to .92 on the *Reading Comprehension* test (n=100); .91 to .94 on the *Reading Vocabulary* test (n=100); and .79 to .91 on the *Listening Comprehension* test (n=170). The difficulty level of each item is reported. Correlation coefficients of the *Progressive Achievement Tests* with other tests of verbal and cognitive skills and with each other are also given. For example, the correlations between the *Reading Comprehension* test and the *Reading Vocabulary* test range from .79 to .86 (three samples were used). The correlation between the *Reading Comprehension* test and the *ACER Reading* for meaning test is .75. There is also considerable data concerning the *Listening Comprehension* test.

References:

Buros, O.K. (Ed.) *Seventh Mental Measurement Yearbook. Vol. II.* Highland Park, New Jersey: The Gryphon Press, 1965.

Elley, W.B. The development of a set of content-referenced tests of reading. Adapted from an address to a seminar at Harcourt Brace Jovanovich, Inc., 1971.

Reading Progress Scale by Ronald P. Carver. © 1970, 1971, Revrac Publications.

Description: The test is designed to measure reading-input performance, the process by which graphic symbols contained in reading material are decoded or translated into a form which can be subsequently understood and stored. The test is not designed to measure reading comprehension. The test results indicate what level of reading material an individual can process. The instrument was constructed on the basis of work done by Bormuth (1969) and Carver (1971). The author states in the manual that "the test is not a traditional norm referenced test. It has not been designed to maximally discriminate between individuals. It is a type of criterion-referenced test." The author suggests that the Reading Progress Scale be used in research, in evaluation studies, and for placing students in reading groups. Two forms are available.

Format and Administration: The test consists of four paragraphs at different difficulty levels, ranging from grades 1 to 3 reading material to grades 10 to 12 (adult) reading material. In each paragraph, an alternative word has been provided for every fifth word; the individual must decide which of the two words fits into the context. The paragraphs were developed via a mechanical procedure and are similar to those developed by means of the cloze technique. The test was pretested on students in grades 3 to 12 but may be useful with younger and older people as well. It may be administered to groups and is timed for seven minutes. No training is necessary to administer the test.

Response Mode and Scoring: The student can put an *X* in the box next to the word he/she feels fits into the sentence or he/she may use a machine-scorable answer sheet; therefore the test is machine and/or manually scorable. Criterion-level scores are given.

Technical Information: Results of pretesting the test on

500 students in grades 3 to 12 are given. The alternate form reliability coefficient is estimated to be .84. Evidence of validity is provided by a comparison of scores on the Reading Progress Scale to teachers' estimates of reading level.

References:

Bormuth, J.R. Development of readability analyses. USOE Final Report Project No. 7-0052, March 1969.

Carver, R.P. A computer model of reading and its implications for measurement and research. *Reading Research Quarterly*, 1971.

SEL/Project Language Level II, Kindergarten © 1971, Southeastern Education Lab, 3450 International Boulevard, Suite 221, Atlanta, Georgia.

Description: This test was developed for use in conjunction with the SEL/Project Language Level II kindergarten program, a 32-lesson set of materials for rural disadvantaged kindergarten children. The lessons are designed to alleviate language deficiencies through school-readiness activities and through cultural and communication experiences. The items on the test are tied to program activities.

Format and Administration: The test, consisting of 28 items, is individually administered. The child responds verbally or through physical activity. The array of props necessary to administer the test includes: texture samples, a xylophone, a flannel board, pictures, and a toy car.

Response Mode and Scoring: The teacher who administers the test records and tallies responses.

Technical Information: None given in ED 055 748.

References:

The curriculum guides are available as ERIC numbers PS 004 669 and PS 005 021.

Territorial Decentration Test by Joseph Stoltman, 1971 (not copyrighted). Available from Mr. Joseph P. Stoltman, Department of Geography, Western Michigan University, Kalamazoo, Michigan.

Description: The framework for this instrument is Piaget's theory of cognitive development. The test is designed to measure the child's territorial decentration—one indication of how far the child has progressed in the transition from a preoperational to a logical mode of thought. There are four subtests: Verbal Territorial Identification; Verbal Territorial Relationship; Territorial Inclusion Using Written Symbols; Territorial Inclusion Using Props.

Format and Administration: The test is individually administered, requiring about 20 minutes. No precise indication of the age group for which the test is appropriate is given, but the vocabulary level is limited to that of "young children." An array of props (outline maps, colored discs and rectangles, and premarked sheets) must be obtained or made by the examiner.

Response Mode and Scoring: The examiner scores the test by following instructions in the manual. Criterion scores for Piaget's levels are given.

Technical Information: No technical information is given.

References: None cited by the author.

Tests of Achievement in Basic Skills—Mathematics by James C. Young, © 1970-1973, Educational and Industrial Testing Service.

Description: The Tests of Achievement in Basic Skills—Mathematics (TABS) is part of the Individualized Mathematics Program (IMP), a system of mathematics instruction based on performance objectives. The objectives are divided into three categories: Arithmetic Skills, Geometry-Measurement-Application, and Modern Concepts. The MATH-PAKS (lesson units directly related to the performance objectives) are eight-page individual instruction booklets containing a diagnostic pretest, examples and drill problems, a practice self test, and a separate posttest. The TABS are used for placing students in levels of the IMP. Since each is available in two parallel forms, they may be used for posttesting as well. The TABS may also be used independently of the IMP and are currently available for a wider range of grade levels than IMP. The tests are criterion-referenced in that "the items were developed to provide an operational assessment of each of the specified educational objectives in IMP."

There are seven TABS-Mathematics levels. They are Kindergarten, Grade 1, Grade 2, Level A for grades 3 and 4, Level B for grades 4 through 6, Level C for grades 7 through 9, and Level D for grades 10 through 12. The *Kindergarten Test* covers: (1) numeration (Arithmetic Skills); (2) recognition of simple geometric shapes, inside-outside, length, and weight (Geometry-Measurement-Application); and (3) problems involving sets, sequences, and the number line (Modern Concepts). The *Grade 1 Test* covers: (1) numeration, addition and subtraction of whole numbers, identification of halves, thirds, and fourths (Arithmetic Skills); (2) geometric shapes, length, time, money, liquid measure, purchasing items, and identifying when to add (Geometry-Measurement-Application); and (3) set concepts, one-to-one correspondence, inequality, expanded notation, sequences, and odd-even concepts (Modern Concepts). The *Grade 2 Test* covers: (1) numeration, number line operations, addition and subtraction of whole numbers, multiplica-

tion with factors not to exceed 5, and the identification of fractional parts (Arithmetic Skills); (2) geometric shapes, length, time, liquid measure, weight, money, and story problems (Geometry-Measurement-Application); and (3) sets, one-to-one correspondence, inequality, expanded notation, and sequences (Modern Concepts). *Level A* covers: (1) addition, subtraction and multiplication of whole numbers, and identification of fractions (Arithmetic Skills); (2) geometric shapes, length, time, money, liquid measure, purchasing items and identifying when to use numerical operations (Geometry-Measurement-Application); and (3) set concepts, one-to-one correspondence, inequality, expanded notation, sequences, and odd-even concepts (Modern Concepts). *Level B* covers: (1) operations with whole and rational numbers (Arithmetic Skills); (2) basic geometric concepts, arithmetic measurements, and application of basic mathematics skills to practical problems (Geometry-Measurement-Application); and (3) sequences, number properties, primes, sets, expanded notations, ordered pairs, and divisibility rules (Modern Concepts). *Level C* covers: (1) operations with integers, rational numbers, irrational numbers, and literal numbers (Arithmetic Skills); (2) basic geometric concepts, arithmetic measurements, and application of basic mathematic skills to practical problems (Geometry-Measurement-Application); and (3) predictions, sequences, functions, number properties, properties of operations, primes, other number bases, and sets (Modern Concepts). *Level D* covers: (1) Arithmetic Skills; and (2) basic geometric concepts, arithmetic measurements, and application of basic mathematic skills to practical problems (Arithmetic Applications).

Format and Administration: Each TABS requires about one hour of the subject's time. The *Kindergarten Test*, which also may be used with preschoolers, is administered by means of an easel flip chart. Higher-level tests utilize a multiple-choice format. No special training is needed to administer the tests.

Response Mode and Scoring: Separate special response booklets are used with the flip chart test; machine-scorable answer sheets accompany the higher level tests. Hand-scoring stencils are available. The EDITS machine-scoring service includes a class profile report as well as individual reports.

Technical Information: Not all levels of the program have complete technical information. Levels B and C (two forms each) have considerable data associated with them. Procedures used for developing these tests are explicitly outlined. Item difficulties, as well as normative data from a national sample, are given for both tests. Three types of reliability estimates were obtained for Level B. Test-retest reliability coefficients ranged from .63-.86; alternate forms from .68 to .87; and K-R 20 from .83 to .98. Level C yielded alternate form coefficients which ranged from .87 to .97; and split-half coefficients which ranged from .91 to .97.

Content validity was assured by examining texts and instructional materials commonly used at each grade level and by soliciting the opinions of subject matter specialists in representative schools.

References:

Young, J.C., Knapp, R.R., & Michael, W.B. The validity of the Tests of Achievement in Basic Skills for predicting achievement in general mathematics and algebra. *Educational and Psychological Measurement*, 1970, 30, 951-954.

Visual Analysis Test (VAT) by Jerome Rosner. 1971 (not copyrighted). Learning Research and Development Center, University of Pittsburgh.

Description: This test measures the ability to copy geometric designs, a predictor of general visual-motor development. The test items can be used as teaching objectives with the expectation that competency acquired in the behaviors these items represent will be generalized to other related tasks. The author has designed a visual-motor curriculum, which contains over 30 behavioral objectives, and he states that for the purpose of this test 27 of these were treated as test items, producing a criterion-referenced test.

Format and Administration: The test is individually administered. It is intended for children in grades K through 2 but may be useful with somewhat younger and older children. There are 27 figures on the test, arranged in order of complexity. No estimate of how much time is necessary to complete the task is given.

Response Mode and Scoring: The examinees copy the figures and the administrator hand-scores the items using a transparent overlay which is provided.

Technical Information: An inter-rater reliability of .98 was estimated with three project staff members scoring independently. The author's initial report (see reference) gives mean scores, medians, and score ranges for grades K through 2 in each of three schools (N=667). Item difficulty was analyzed to confirm the hierarchy of behavioral objectives as represented by the 27 items. Scores on the VAT were correlated with scores on the Rutgers Drawing Test, Form A (N=470). A correlation coefficient (N=197) was also calculated for those in second grade who took both the VAT and Form B of the Rutgers Drawing Test. In both instances, the correlation coefficients were positive and high (r=.80 with Form A; r=.68 with Form B).

Reference:

Rosner, J. The visual analysis test: An initial report. University of Pittsburgh, Learning R & D Center, 1971.

Wisconsin Tests of Reading Skill Development: Word Attack: Developmental Edition by Wayne Otto. © 1970, National Computer Systems, Inc./Interpretive Scoring Systems.

Description: This series of word attack tests was developed for use with the Wisconsin Design for Reading Skill Development, a prototype for an individually guided reading skill program for elementary schools. The curriculum developers claim that the Wisconsin design "represents a systematic attempt to assess individual pupils' skill development status by means of criterion-referenced tests with respect to explicitly stated behaviors related to each skill." Word attack is only one of six skill areas covered in the design. Of the word attack behavioral objectives, some are assessed through formal criterion-referenced tests and others are felt to lend themselves better to informal assessment through Guides to Informal Individual Skill Observation (available separately). There are four Word Attack tests. *Level A* consists of subtests on rhyming words, rhyming phrases, shapes, letters and numbers, words and phrases, initial consonants, distinguishing colors. *Level B* has subtests covering sight vocabulary, beginning consonant sounds, ending consonant sounds, consonant blends, rhyming elements, short vowels, consonant digraphs, compound words, contractions, base words and endings, plurals, and possessives. *Level C* includes subtests on the following: sight vocabulary; consonants and their variant sounds; consonant blends, long vowel sounds, vowel + r, a + l, a + w, diphthongs; long and short oo; consonant digraphs; base words; plurals; homonyms; synonyms and antonyms; and multiple meanings. *Level D* covers sight vocabulary, three-letter consonant blends, silent letters, syllabication, accent, schwa, and possessives.

Format and Administration: These multiple-choice tests are administered to small groups of children. Level A, appropriate for the end of kindergarten or the beginning of first grade, requires four sittings of 20-30 minutes each. Level B, appropriate for the end of first grade or the beginning of second grade, also requires four sittings of about 20 minutes. Level C, appropriate for the end of grade 2 or the beginning of grade 3, requires four sittings of 30-40 minutes each. Level D, appropriate for the end of grade 3 or the beginning of grade 4, may be administered in two 35 to 40 minute sittings.

Response Mode and Scoring: The students use a special answer sheet. The tests can be machine-scored by National Computer Systems, Inc., or they can be hand-scored.

Technical Information: There are no group norms. The reliabilities of subtests obtained during the development of the test range from .70s to .90s.

References:

Askov, E.N. The Word Attack element of the Wisconsin design. Paper presented at the National Reading Conference, Tampa, Florida, December, 1971.

Fischback, J.J. Study of relationships of reading mastery level to general reading achievement to validate diagnostic reading tests. Paper presented at Annual Meeting of American Educational Research Association, New York, 1971.

Otto, W., & Askov, E.N. The Wisconsin design for reading skill development. Paper presented at meeting of International Reading Association, Atlantic City, New Jersey, April 1971.

Rude, R.T. Implementation and field testing of the Wisconsin design. Paper presented at National Reading Conference, Tampa, Florida, December 1971.

Woodcock Reading Mastery Tests by Richard W. Woodcock. © 1973, American Guidance Service Inc.

Description: This is a battery of five reading skills tests: Letter Identification; Word Identification; Word Attack; Word Comprehension; and Passage Comprehension. Scores on the five tests yield diagnostic and instructional information and may be used to predict the student's potential for success at selected reading tasks. Two forms, A and B, are available. The manual for the battery provides reference scales whereby teachers can obtain a criterion-referenced interpretation of a student's performance.

Format and Administration: The tests are individually administered and require a total of 20 to 30 minutes. Items are presented on cards set up in a book. The examiner selects from the kit those items which are appropriate to the student's level. The tests may be used for grades K through 12. No special training is required of the examiner.

Scoring: The test is scored by the examiner. Raw scores may be converted to percentile ranks, grade-level scores, age scores, or standard scores. Complete guidelines for interpreting results are provided.

Technical Information: The test items were analyzed, calibrated, and selected by means of Rasch-Wright procedures. In addition to usual norms, separate norms for boys and girls and norms adjusted for socioeconomic status are available. Split-half reliabilities for grades 1, 4, and 10 are in the .90 to .99 range for subtests and in the .97 to .99 range for the total battery. Procedures for assuring content validity are outlined in the manual. A multimethod-multitrait analysis using forms A and B was conducted on data obtained from second- and seventh-grade subjects. This consisted simply of an intercorrelation matrix of data from both forms and sub-

tests. The coefficients ranged from .84 to .94, showing that parallel forms of the test were highly correlated.

References:

Woodcock, R.W. The Peabody-Chicago-Detroit Reading Project—a report of the second-year results. In J.R. Block (Ed.). *i.t.a. as a language arts medium.* Hempstead, N.Y.: The i.t.a. Foundation, Hofstra University, 1968, 188-196.

Woodcock, R.W. Vocabulary analysis of seven basic reading series. AGS Paper No. 7. Circle Pines, Minn.: American Guidance Service, 1971. (Available from AGS)

Woodcock, R.W., & Dahl, M.N. A common scale for the measurement of person ability and test item difficulty. AGS Paper No. 10. Circle Pines, Minn.: American Guidance Service, 1971.

REFERENCES*

Block, J.H. (Ed.) *Mastery learning: theory and practice.* New York: Holt, Rinehart and Winston, 1971.

Brazziel, W.F. Criterion-referenced tests—some trends and prospects. *Today's Education.* 1972, 61, 52-53.

Cox, R.C. Evaluative aspects of criterion-referenced measures. Paper presented at Annual AERA Meeting, Minneapolis, 1970. ED 038 679.

Ebel, R.L. Criterion-referenced measurements: Limitations. *School Review.* 1971, 79, 282-288.

Fitzgibbon, T.J. Norm referenced and criterion-referenced tests from a publishers' point of view. Paper presented at annual meeting of American Psychological Association, Honolulu, Hawaii, 1972. ED 069 787.

Fremer, J. Criterion-referenced interpretations of survey achievement tests. TM-72-1. Princeton, New Jersey: Educational Testing Service, 1972. ED 065 533.

Garvin, A.D. The applicability of criterion-referenced measurement by content area and level. Paper presented at Annual AERA Meeting, Minneapolis, 1970. ED 041 038.

Gentile, J.R. Toward excellence in teaching: Grading practices, 1971 ED 061 264.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, 18, 519-521.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In Robert L. Thorndike (Ed.). *Educational Measurement.* Washington, D.C.: American Council on Education, 1971.

Hawes, G.R. Criterion-referenced testing: No more losers, no more norms, no more parents raising storms. *Nation's Schools,* 1973, 91, 35-41.

Housden, J.L., & LeGear, L. An emerging model: Criterion-referenced evaluation. (in press)

Hull, C.L. The place of innate individual differences in a natural-science theory of behavior. *Psychological Review,* 1945, 52, 55-60.

Ivens, S.H. A pragmatic approach to criterion-referenced measures. Paper presented to National Council on Measurement in Education, Chicago, April 1972. ED 064 406.

Jackson, R. Developing criterion-referenced tests. ERIC/TM Report 1, 1970. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement and Evaluation, 1970. ED 041 052.

Klein, S. Ongoing evaluation of educational programs. Paper presented at annual meeting of the American Psychological Association, Honolulu, Hawaii, September 1972. ED 069 725.

Klein, S.P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC/TM Report 26, 1973. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1973. ED 083 284.

Kriewall, T.E. Aspects and applications of criterion-referenced tests. Paper presented at Annual AERA Meeting, Chicago, 1972. ED 063 333.

Livingston, S.A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement.* 1972, 9, 13-25.

Shub, A.N. Criterion-referenced measurement: Proven ideas. Paper presented at annual meeting of National Council on Measurement in Education, New Orleans, 1973.

Skinner, B.F. *The behavior of organisms: An experimental analysis.* New York: Appleton-Century Crofts, 1938.

*Items followed by an ED number (for example ED 069 762) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of *Resources in Education* for the address and ordering information.